Multivariate mixtures of Erlangs for density estimation under censoring

Roel Verbelen*[1], Katrien Antonio[1,2], and Gerda Claeskens[1]

[1]LStat, Faculty of Economics and Business, KU Leuven, Belgium.
[2]Faculty of Economics and Business, University of Amsterdam, The Netherlands.

August 20, 2015

Abstract

Multivariate mixtures of Erlang distributions form a versatile, yet analytically tractable, class of distributions making them suitable for multivariate density estimation. We present a flexible and effective fitting procedure for multivariate mixtures of Erlangs, which iteratively uses the EM algorithm, by introducing a computationally efficient initialization and adjustment strategy for the shape parameter vectors. We furthermore extend the EM algorithm for multivariate mixtures of Erlangs to be able to deal with randomly censored and fixed truncated data. The effectiveness of the proposed algorithm is demonstrated on simulated as well as real data sets.

Keywords: Multivariate mixtures of Erlangs with a common scale parameter; Density estimation; Censored data; Expectation-maximization algorithm; Maximum likelihood.

1 Introduction

We present an estimation technique for fitting multivariate mixtures of Erlang distributions (MME). We suggest an efficient initialization method and adjustment strategy for the values of the shape parameter vectors of an MME, which has been underexposed in the literature. The fitting procedure is also extended to take random censoring and fixed truncation into account.

*Corresponding author. E-mail adress: roel.verbelen@kuleuven.be

The proposed algorithm has been implemented in `R` and is available online at `www.http://feb.kuleuven.be/roel.verbelen`. Data are censored in case you only observe an interval in which a data point is lying without knowing its exact value. Truncation entails that it is only possible to observe the data of which the values lie in a certain range. Censoring and/or truncation is often the case in applications such as loss modeling (finance and actuarial science), clinical experiments (survival/failure time analysis), veterinary studies (e.g. mastitis studies), and duration data (econometric studies).

The class of MME is introduced by Lee and Lin (2012). MME form a highly flexible class of distributions as they are dense in the space of positive continuous multivariate distributions in the sense of weak convergence, extending this property of the univariate class (Tijms, 1994). An overview of the analytical and distributional properties of mixtures of Erlangs can be found in Klugman et al. (2013), Willmot and Lin (2011) and Willmot and Woo (2007). Parameter estimation in the univariate case is treated in Lee and Lin (2010) and extended to be able to deal with randomly censored and fixed truncated data in Verbelen et al. (2015).

Mixtures of Erlangs have received most attention in the field of actuarial science. Cossette et al. (2013a) model the joint distribution of a portfolio of dependent risks using univariate mixtures of Erlangs as marginals along with the Farlie-Gumbel-Morgenstern (FGM) copula. Cossette et al. (2013b) and Mailhot (2012) study the bivariate lower and upper orthant Value-at-Risk and use MME as an illustration. Willmot and Woo (2015) study the analytical properties of the MME class. They motivate the use of MME in actuarial science and illustrate how their tractability leads to closed-form expressions.

The use of MME should be regarded as a multivariate density estimation technique, not as as a type of model-based clustering. The MME model can be seen as semiparametric, since the mixture components have a specific parametric form, whereas the mixing weights can have a nonparametric nature, and is an interesting alternative to the use of copulas, which is the dominant choice to model multivariate data in a two stage procedure, separating the dependence structure from the marginal distributions (see e.g. Joe, 1997; Nelsen, 2006). In contrast, MME are able to model the multivariate data directly on the original scale. The MME model enjoys many desirable properties of a multivariate model as listed by Joe (1997, p. 84), see Lee and Lin (2012), with regard to interpretability, closure, flexibility and wide range of dependence, and closed-form representation, often not satisfied for the commonly used copula structures.

An extensive literature exists on mixtures of multivariate normals (see e.g. McLachlan and

Peel, 2001). Lee and Scott (2012) discuss the estimation of multivariate Gaussian mixtures in case the data can be randomly censored and fixed truncated. Due to the limitations of Gaussian mixtures, such as the difficulty in modeling skewed data, non-Gaussian approaches have received an increasing interest over the last years. Important examples include mixtures of multivariate t-distributions (see e.g. Peel and McLachlan, 2000), mixtures of multivariate skew-normal distributions (see e.g. Lin, 2009), and mixtures of multivariate skew-t distributions (see e.g. Lee and McLachlan, 2014). All of these mixture models involve modeling real-valued multivariate random variables, whereas in this paper we consider multivariate positive-valued random variables.

Lee and Lin (2012) show in Theorem 2.3 that a finite multivariate Erlang mixture is a multivariate phase-type distribution, a generalization of the class of univariate phase-type distributions introduced by Assaf et al. (1984). Parameter estimation for phase-type distributions in the bivariate case (Eisele, 2005; Zadeh and Bilodeau, 2013), as in the univariate case (Asmussen et al., 1996; Olsson, 1996), uses the expectation-maximization (EM) algorithm, first introduced by Dempster et al. (1977)

The EM algorithm forms the key to fit an MME to multivariate positive data. Taking censoring and truncation into account when calibrating data using copulas is cumbersome, especially in more than two dimensions, due to complicated forms of the likelihood (see e.g. Georges et al., 2001) which are hard to optimize numerically. This is, as we will show, not the case for the MME class due to the EM algorithm. As opposed to the traditional way of dealing with grouped and truncated data using the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2007, p. 66; McLachlan and Peel, 2001, p. 257; McLachlan and Jones, 1988), we follow the approach of Lee and Scott (2012), as was done in the univariate setting (Verbelen et al., 2015).

We demonstrate the effectiveness of our proposed algorithm and the practical use of MME on a simulated dataset, the old faithful geyser data and a four-dimensional dataset of interval and right censored udder quarter infection times, each time highlighting one of the analytical aspects of MME.

2 Multivariate Erlang mixtures with a common scale parameter

In this section, we briefly revise the definition of a multivariate mixture of Erlang distributions with a common scale parameter and the denseness property of this distributional class. These

formulas are extended in Section 3.1 and 3.2 towards censoring and truncation.

The Erlang distribution is a positive continuous distribution with density function

$$f(x;r,\theta) = \frac{x^{r-1}e^{-x/\theta}}{\theta^r(r-1)!} \qquad \text{for } x > 0\,, \tag{1}$$

where r, a positive integer, is the shape parameter and $\theta > 0$ the scale parameter (the inverse $\lambda = 1/\theta$ is called the rate parameter). The cumulative distribution function is obtained by integrating (1) by parts r times

$$F(x;r,\theta) = 1 - \sum_{n=0}^{r-1} e^{-x/\theta}\frac{(x/\theta)^n}{n!} = \frac{\gamma(r,x/\theta)}{(r-1)!}\,, \tag{2}$$

using the lower incomplete gamma function defined as $\gamma(s,x) = \int_0^x z^{s-1}e^{-z}dz$.

A univariate Erlang distribution is in fact a gamma distribution of which the shape parameter is a positive integer and can therefore be seen as the distribution of a sum of i.i.d. exponential random variables. Lee and Lin (2012) define a d-variate Erlang mixture as a mixture such that each mixture component is the joint distribution of d independent Erlang distributions with a common scale parameter $\theta > 0$. The dependence structure is captured by the combination of the positive integer shape parameters of the Erlangs in each dimension. We denote the positive integer shape parameters of the jointly independent Erlang distributions in a mixture component by the vector $\boldsymbol{r} = (r_1,\ldots,r_d)$ and the set of all shape vectors with non-zero weight by $\mathcal{R}$. The mixture weights are denoted by $\boldsymbol{\alpha} = \{\alpha_{\boldsymbol{r}}\,|\boldsymbol{r} \in \mathcal{R}\}$ and must satisfy $\alpha_{\boldsymbol{r}} \geqslant 0$ and $\sum_{\boldsymbol{r}\in\mathcal{R}} \alpha_{\boldsymbol{r}} = 1$. The density of a d-variate Erlang mixture evaluated in $\boldsymbol{x} = (x_1,\ldots,x_d)$ with $x_j > 0$ for $j = 1,\ldots,d$ can then be written as

$$f(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{r},\theta) = \sum_{\boldsymbol{r}\in\mathcal{R}} \alpha_{\boldsymbol{r}} f(\boldsymbol{x};\boldsymbol{r},\theta) = \sum_{\boldsymbol{r}\in\mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} f(x_j;r_j,\theta) = \sum_{\boldsymbol{r}\in\mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} \frac{x_j^{r_j-1}e^{-x_j/\theta}}{\theta^{r_j}(r_j-1)!}\,. \tag{3}$$

The following property states that for any positive multivariate distribution there exists a sequence of multivariate Erlang distributions that weakly converges to the target distribution. The proof is given in the appendix of Lee and Lin (2012).

Property 1 (Lee and Lin 2012)**.** *The class of multivariate Erlang mixtures of form (3) is dense in the space of positive continuous multivariate distributions in the sense of weak convergence. More specifically, let $f(\boldsymbol{x})$ be the density function of a d-variate positive random variable with*

cumulative distribution function $F(\boldsymbol{x})$. *For any given* $\theta > 0$, *define the following* d*-variate Erlang mixture*

$$f(\boldsymbol{x};\theta) = \sum_{r_1=1}^{\infty} \cdots \sum_{r_d=1}^{\infty} \alpha_{\boldsymbol{r}}(\theta) \prod_{j=1}^{d} f(x_j; r_j, \theta)\,, \tag{4}$$

with mixing weights

$$\alpha_{\boldsymbol{r}}(\theta) = \int_{(r_1-1)\theta}^{r_1\theta} \cdots \int_{(r_d-1)\theta}^{r_d\theta} f(\boldsymbol{x}) d\boldsymbol{x}\,. \tag{5}$$

Then $\lim_{\theta\to 0} F(\boldsymbol{x};\theta) = F(\boldsymbol{x})$ *for each point* $\boldsymbol{x}$ *at which* F *is continuous.*

In Property 1, for any given common scale $\theta > 0$, an infinite multivariate mixture of Erlangs in (4) is considered using combinations of shapes from 1 to infinity in each marginal dimension. The weights in (5) of the components in the mixture are defined by integrating the density over the corresponding d-dimensional rectangle of the d-dimensional grid formed by the shape parameters multiplied with the common scale. When the value of the common scale θ decreases, this grid becomes more refined and the sequence of Erlang mixtures converges to the underlying cumulative distribution function.

Next to its flexibility, Lee and Lin (2012) show that it is easy to work analytically with this class of distributions due to the independence structure of the Erlang distributions within each mixture component. This leads to explicit expressions of many distributional quantities such as the characteristic function, the joint moments and bivariate measures of association (Kendall's tau and Spearman's rho). The authors further reveal interesting closure properties, such as the fact that each p-variate marginal or conditional distribution with $p \leqslant d$ can again be written as a p-variate Erlang distribution. The same property holds for the distribution of the multivariate excess losses (actuarial science context) or multivariate residual lifetimes (survival analysis context). Furthermore, the distribution of the sum of the component random variables of an MME distributed random variable is a univariate Erlang mixture distribution.

Willmot and Woo (2015) consider an extension of the MME class, allowing different scale parameters in each dimension. However, in Proposition 1 they show how a multivariate mixture of Erlangs distribution with different scale parameters can be rewritten as a multivariate mixture of Erlangs distribution with a common scale parameter, which is smaller than all original scales. We thus concentrate on models with a common scale parameter.

3 Parameter estimation

The parameters of an MME to be estimated are the common scale parameter θ, the mixture weights $\boldsymbol{\alpha} = \{\alpha_{\boldsymbol{r}} \,|\boldsymbol{r} \in \mathcal{R}\}$ and the set of corresponding shape parameter vectors $\mathcal{R}$. Lee and Lin (2012) propose an EM algorithm in order to find the maximum likelihood estimators for $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \theta)$, given a fixed set of shape parameter vectors $\mathcal{R}$. Model selection for the number of mixture components and the corresponding values of the shape parameter vectors is based on an information criterion, similar to the univariate strategy of Lee and Lin (2010) and Verbelen et al. (2015).

The two main novelties we present in this paper are (i) an extension of the EM algorithm to be able to deal with randomly censored and fixed truncated data and (ii) a computationally more efficient initialization and adjustment strategy for the shape parameter vectors in order to make the estimation procedure more flexible and effective. The improvements (i) and (ii) allow us to analyze realistic data with diverse forms of dependence in contrast to the simulated example in Lee and Lin (2012) with a simple structure.

First we discuss how we represent a censored and truncated sample and evaluate the expression of the likelihood. The form of the complete data log-likelihood is given next, followed by the adjusted EM algorithm and a discussion on some asymptotic properties. In Section 4, we present the initialization and selection of the shape parameter vectors.

3.1 Randomly censored and fixed truncated data

We represent a censored sample, truncated to the fixed range $[\boldsymbol{t}^l, \boldsymbol{t}^u]$, by $\mathcal{X} = \{\,(\boldsymbol{l}_i, \boldsymbol{u}_i) |\, i = 1, \ldots, n\}$. The lower and upper truncation points are $\boldsymbol{t}^l = (t_1^l, \ldots, t_d^l)$ and $\boldsymbol{t}^u = (t_1^u, \ldots, t_d^u)$, which are common to each observation $i = 1, \ldots, n$. The lower and upper censoring points are $\boldsymbol{l}_i = (l_{i1}, \ldots, l_{id})$ and $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})$. It holds that $\boldsymbol{t}^l \leqslant \boldsymbol{l}_i \leqslant \boldsymbol{u}_i \leqslant \boldsymbol{t}^u$ for $i = 1, \ldots, n$. $t_j^l = 0$ and $t_j^u = \infty$ mean no truncation from below and above for the jth dimension, respectively. The censoring status for the jth dimension of observation i is determined as follows:

$$
\begin{array}{ll}
\text{Uncensored:} & t_j^l \leqslant l_{ij} = u_{ij} =: x_{ij} \leqslant t_j^u \\
\text{Left Censored:} & t_j^l = l_{ij} < u_{ij} < t_j^u \\
\text{Right Censored:} & t_j^l < l_{ij} < u_{ij} = t_j^u \\
\text{Interval Censored:} & t_j^l < l_{ij} < u_{ij} < t_j^u.
\end{array}
$$

Thus, l_{ij} and u_{ij} should be interpreted as the lower and upper endpoints of the interval that

contains the jth element of observation i. A missing value in dimension j for observation i can also be dealt with by setting $l_{ij} = t_j^l$ and $u_{ij} = t_j^u$, i.e. treating the missing value as a data point being interval censored between the lower and upper truncation points.

The likelihood of a censored and truncated sample of a multivariate Erlang distribution is given by

$$\mathcal{L}(\boldsymbol{\Theta}; \mathcal{X}) = \prod_{i=1}^{n} \frac{\sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} f(l_{ij}, u_{ij}; r_j, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})}$$

with

$$f(l_{ij}, u_{ij}; r_j, \theta) = \begin{cases} f(x_{ij}; r_j, \theta) & \text{if } l_{ij} = u_{ij} = x_{ij} \\ F(u_{ij}; r_j, \theta) - F(l_{ij}; r_j, \theta) & \text{if } l_{ij} < u_{ij}\,, \end{cases}$$

and

$$\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta}) = \sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} \left[F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta) \right] .$$

The corresponding log-likelihood is

$$l(\boldsymbol{\Theta}; \mathcal{X}) = \sum_{i=1}^{n} \ln \left(\sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} f(l_{ij}, u_{ij}; r_j, \theta) \right) - n \ln \left(\sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} \left[F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta) \right] \right) . \tag{6}$$

This expression is however not workable as it involves the logarithm of a sum and cannot be used to easily find the maximum likelihood estimators for $\boldsymbol{\Theta}$ for a fixed set of positive integer shape parameters $\mathcal{R}$.

3.2 Construction of the complete data likelihood

For an uncensored observation $\boldsymbol{x}_i$, truncated to $[\boldsymbol{t}^l, \boldsymbol{t}^u]$, the probability density function can be rewritten as a mixture

$$\begin{aligned} f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{\Theta}) &= \frac{f(\boldsymbol{x}_i; \boldsymbol{\Theta})}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})} = \frac{\sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} f(x_{ij}; r_j, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})} \\ &= \sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \cdot \frac{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{r}, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})} \cdot \frac{\prod_{j=1}^{d} f(x_{ij}; r_j, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{r}, \theta)} = \sum_{\boldsymbol{r} \in \mathcal{R}} \beta_{\boldsymbol{r}} \cdot f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta)\,, \end{aligned}$$

for $\boldsymbol{t}^l \leqslant \boldsymbol{x}_i \leqslant \boldsymbol{t}^u$ and zero otherwise. The mixing weights $\beta_{\boldsymbol{r}}$ and component density functions are given by, respectively,

$$\beta_{\boldsymbol{r}} = \alpha_{\boldsymbol{r}} \cdot \frac{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{r}, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})} = \alpha_{\boldsymbol{r}} \cdot \frac{\prod_{j=1}^{d} \left[F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta) \right]}{\sum_{\boldsymbol{m} \in \mathcal{R}} \alpha_{\boldsymbol{m}} \prod_{j=1}^{d} \left[F(t_j^u; m_j, \theta) - F(t_j^l; m_j, \theta) \right]} \tag{7}$$

and

$$f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta) = \frac{\prod_{j=1}^{d} f(x_{ij}; r_j, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{r}, \theta)} = \prod_{j=1}^{d} \frac{f(x_{ij}; r_j, \theta)}{F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta)} \,. \tag{8}$$

The weights $\beta_{\boldsymbol{r}}$ are re-weighted versions of the original weights $\alpha_{\boldsymbol{r}}$ by means of the probabilities of the corresponding mixture component to lie in the d-dimensional truncation interval. The component density functions $f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta)$ are truncated versions of the original component density functions $f(\boldsymbol{x}_i; \boldsymbol{r}, \theta)$.

The EM algorithm forms the solution to fit this finite mixture to the censored and truncated data. The idea is to regard the censored sample $\mathcal{X}$ as being incomplete since the uncensored observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ and their associated component-indicators $\boldsymbol{z}_i = \{z_{i\boldsymbol{r}} \,|\boldsymbol{r} \in \mathcal{R}\}$ with

$$z_{i\boldsymbol{r}} = \begin{cases} 1 & \text{if observation } \boldsymbol{x}_i \text{ comes from the mixture component (8)} \\ & \text{corresponding to the shape parameter vector } \boldsymbol{r} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

for $i = 1, \ldots, n$ and $\boldsymbol{r} \in \mathcal{R}$, are not available. The complete data vector, $\mathcal{Y} = \{(\boldsymbol{x}_i, \boldsymbol{z_i}) | i = 1, \ldots, n\}$, contains all uncensored observations $\boldsymbol{x}_i$ and their corresponding mixing component indicator $\boldsymbol{z}_i$. The log-likelihood of the complete sample $\mathcal{Y}$ can then be written as

$$l(\boldsymbol{\Theta}; \mathcal{Y}) = \sum_{i=1}^{n} \sum_{\boldsymbol{r} \in \mathcal{R}} z_{i\boldsymbol{r}} \ln \left(\beta_{\boldsymbol{r}} f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta) \right) \,. \tag{10}$$

3.3 The EM algorithm for censored and truncated data

The EM algorithm finds the maximum likelihood estimators for $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \theta)$, given a fixed set $\mathcal{R}$ of positive integer shape parameter vectors, based on a (possibly) censored and truncated sample by iteratively repeating the following two steps.

E-step Conditional on the incomplete data $\mathcal{X}$ and using the current estimate $\boldsymbol{\Theta}^{(k-1)}$ for $\boldsymbol{\Theta}$, we compute the expectation of the complete log-likelihood (10) in the kth iteration of the E-step:

$$\begin{aligned}
Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)}) &= E(l(\boldsymbol{\Theta};\mathcal{Y}) \mid \mathcal{X};\boldsymbol{\Theta}^{(k-1)}) \\
&= \sum_{i=1}^{n} E\left[\sum_{\boldsymbol{r}\in\mathcal{R}} Z_{i\boldsymbol{r}} \ln\left(\beta_{\boldsymbol{r}} f(\boldsymbol{X}_i;\boldsymbol{t}^l,\boldsymbol{t}^u,\boldsymbol{r},\theta)\right)\middle| \boldsymbol{l}_i,\boldsymbol{u}_i,\boldsymbol{t}^l,\boldsymbol{t}^u;\boldsymbol{\Theta}^{(k-1)}\right] \\
&= \sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} E\left[\ln\left(\beta_{\boldsymbol{r}} f(\boldsymbol{X}_i;\boldsymbol{t}^l,\boldsymbol{t}^u,\boldsymbol{r},\theta)\right)\middle| Z_{i\boldsymbol{r}}=1,\boldsymbol{l}_i,\boldsymbol{u}_i,\boldsymbol{t}^l,\boldsymbol{t}^u;\theta^{(k-1)}\right] \\
&= \sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \Bigg[\ln(\beta_{\boldsymbol{r}}) + \sum_{j=1}^{d}(r_j-1)E\left(\ln(X_{ij})\middle| Z_{i\boldsymbol{r}}=1,l_{ij},u_{ij},t_j^l,t_j^u;\theta^{(k-1)}\right) \\
&\quad - \frac{1}{\theta}\sum_{j=1}^{d} E\left(X_{ij}\middle| Z_{i\boldsymbol{r}}=1,l_{ij},u_{ij},t_j^l,t_j^u;\theta^{(k-1)}\right) - \sum_{j=1}^{d} r_j\ln(\theta) - \sum_{j=1}^{d}\ln((r_j-1)!) \\
&\quad - \sum_{j=1}^{d}\ln\left(F(t_j^u;r_j,\theta)-F(t_j^l;r_j,\theta)\right)\Bigg] . \qquad (11)
\end{aligned}$$

In the fourth equality, we apply the law of total expectation and denote the posterior probability that observation i belongs to the mixture component corresponding to the shape parameters $\boldsymbol{r}$ as $z_{i\boldsymbol{r}}^{(k)}$. These posterior probabilities can be computed using Bayes' rule,

$$\begin{aligned}
z_{i\boldsymbol{r}}^{(k)} &= P(Z_{i\boldsymbol{r}}=1 \mid \boldsymbol{l}_i,\boldsymbol{u}_i,\boldsymbol{t}^l,\boldsymbol{t}^u;\boldsymbol{\Theta}^{(k-1)}) \\
&= \frac{\beta_{\boldsymbol{r}}^{(k-1)}\prod_{j=1}^{d}\left[f(l_{ij},u_{ij};r_j,\theta^{(k-1)})\Big/\left(F(t_j^u;r_j,\theta^{(k-1)})-F(t_j^l;r_j,\theta^{(k-1)})\right)\right]}{\sum_{\boldsymbol{m}\in\mathcal{R}}\beta_{\boldsymbol{m}}^{(k-1)}\prod_{j=1}^{d}\left[f(l_{ij},u_{ij};m_j,\theta^{(k-1)})\Big/\left(F(t_j^u;m_j,\theta^{(k-1)})-F(t_j^l;m_j,\theta^{(k-1)})\right)\right]} \\
&= \frac{\alpha_{\boldsymbol{r}}^{(k-1)}\prod_{j=1}^{d} f(l_{ij},u_{ij};r_j,\theta^{(k-1)})}{\sum_{\boldsymbol{m}\in\mathcal{R}}\alpha_{\boldsymbol{m}}^{(k-1)}\prod_{j=1}^{d} f(l_{ij},u_{ij};m_j,\theta^{(k-1)})} . \qquad (12)
\end{aligned}$$

using (7), for $i=1,\ldots,n$ and $\boldsymbol{r}\in\mathcal{R}$.

Since the terms in (11) for $Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)})$ containing $E\left(\ln(X_{ij})\middle| Z_{i\boldsymbol{r}}=1,l_{ij},u_{ij},t_j^l,t_j^u;\theta^{(k-1)}\right)$ do not depend on the unknown parameter vector $\boldsymbol{\Theta}$, they will not play a role in the EM algorithm. In the E-step, we need to compute the expected value of X_{ij} conditional on the censoring and truncation points and the mixing component $Z_{i\boldsymbol{r}}$ for the current value $\boldsymbol{\Theta}^{(k-1)}$ of the parameter vector. For $i=1,\ldots,n$ and $\boldsymbol{r}\in\mathcal{R}$, we have

$$E\left(X_{ij}\middle| Z_{i\boldsymbol{r}}=1,l_{ij},u_{ij},t_j^l,t_j^u;\theta^{(k-1)}\right)$$

$$
\begin{aligned}
&= \int_{l_{ij}}^{u_{ij}} x \frac{f(x; r_j, \theta^{(k-1)})}{F(u_{ij}; r_j, \theta^{(k-1)}) - F(l_{ij}; r_j, \theta^{(k-1)})} dx \\
&= \frac{r_j \theta^{(k-1)}}{F(u_{ij}; r_j, \theta^{(k-1)}) - F(l_{ij}; r_j, \theta^{(k-1)})} \int_{l_{ij}}^{u_{ij}} \frac{x^{r_j} e^{-x/\theta^{(k-1)}}}{\left(\theta^{(k-1)}\right)^{r_j+1} r_j!} dx \\
&= \frac{r_j \theta^{(k-1)} \left(F(u_{ij}; r_j + 1, \theta^{(k-1)}) - F(l_{ij}; r_j + 1, \theta^{(k-1)})\right)}{F(u_{ij}; r_j, \theta^{(k-1)}) - F(l_{ij}; r_j, \theta^{(k-1)})} ,
\end{aligned} \tag{13}
$$

in case $l_{ij} < u_{ij}$ and in case $l_{ij} = u_{ij} = x_{ij}$, the observation is uncensored and the expression is equal to x_{ij}.

M-step In the kth iteration of the M-step, we maximize the expected value (11) of the complete data log-likelihood obtained in the E-step with respect to the parameter vector $\boldsymbol{\Theta}$ over all $(\boldsymbol{\beta}, \theta)$ with $\beta_{\boldsymbol{r}} \geqslant 0$, $\sum_{\boldsymbol{r} \in \mathcal{R}} \beta_{\boldsymbol{r}} = 1$ and $\theta > 0$. The maximization with respect to the mixing weights $\boldsymbol{\beta}$, requires the maximization of

$$
\sum_{i=1}^{n} \sum_{\boldsymbol{r} \in \mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \ln(\beta_{\boldsymbol{r}}) ,
$$

which can be done analogously as in the univariate case, yielding

$$
\beta_{\boldsymbol{r}}^{(k)} = n^{-1} \sum_{i=1}^{n} z_{i\boldsymbol{r}}^{(k)} \qquad \text{for } \boldsymbol{r} \in \mathcal{R} . \tag{14}
$$

The average over the posterior probabilities of belonging to the jth component in the mixture forms the new estimator for the prior probability β_j in the truncated mixture.

We set the first order partial derivative with respect to θ equal to zero in order to maximize $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ over θ (see Appendix A), leading to the following M-step equation:

$$
\theta^{(k)} = \frac{n^{-1} \sum_{i=1}^{n} \sum_{\boldsymbol{r} \in \mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} E\left(X_{ij} \middle| Z_{i\boldsymbol{r}} = 1, l_{ij}, u_{ij}, t_j^l, t_j^u; \theta^{(k-1)}\right) - T^{(k)}}{\sum_{\boldsymbol{r} \in \mathcal{R}} \beta_{\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} r_j} , \tag{15}
$$

with

$$
T^{(k)} = \sum_{\boldsymbol{r} \in \mathcal{R}} \beta_{\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} \left. \frac{\left(t_j^l\right)^{r_j} e^{-t_j^l/\theta} - \left(t_j^u\right)^{r_j} e^{-t_j^u/\theta}}{\theta^{r_j - 1} (r_j - 1)! \left(F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta)\right)} \right|_{\theta = \theta^{(k)}} .
$$

Similar to the univariate case (Verbelen et al., 2015), the new estimator $\theta^{(k)}$ in (15) for the common scale parameter θ has the interpretation of the expected total mean divided by the weighted total shape parameter in the mixture minus a correction term $T^{(k)}$ due to the truncation. Since

$T^{(k)}$ in (15) depends on $\theta^{(k)}$ and has a complicated form, it is not possible to find an analytical solution. Therefore, we use a Newton-type algorithm, with the previous value of θ, i.e. $\theta^{(k-1)}$, as starting value, to solve the equation.

We iterate the E- and M-step until the difference in log-likelihood $l(\boldsymbol{\Theta}^{(k)}; \mathcal{X}) - l(\boldsymbol{\Theta}^{(k-1)}; \mathcal{X})$ between two iterations becomes sufficiently small. By inverting expression (7), we retrieve the maximum likelihood estimator of the original mixing weights $\alpha_{\boldsymbol{r}}^{(k)}$ for $\boldsymbol{r} \in \mathcal{R}$. We first compute

$$\widetilde{\alpha}_{\boldsymbol{r}} = \frac{\widehat{\beta_{\boldsymbol{r}}}}{\prod_{j=1}^{d} \left[F(t_j^u; r_j, \widehat{\theta}) - F(t_j^l; r_j, \widehat{\theta}) \right]} \qquad \text{for } \boldsymbol{r} \in \mathcal{R}\,, \tag{16}$$

where $\widehat{\beta_{\boldsymbol{r}}}$ and $\widehat{\theta}$ denote the values in the final EM step, and then normalize the weights such that they sum to 1.

Using the EM algorithm, the log likelihood (6) increases with each iteration (McLachlan and Krishnan, 2007). The estimator for $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \theta)$ obtained from the EM algorithm has the same limit as the maximum likelihood estimator, whenever the starting value is adequately chosen. Hence, the maximum likelihood asymptotic theory in terms of consistency, asymptotic normality and asymptotic efficiency applies. Within the EM framework, the asymptotic covariance matrix of the maximum likelihood estimator can be assessed (McLachlan and Krishnan, 2007).

These asymptotic results can only be applied with respect to $\boldsymbol{\Theta}$, given a fixed shape set $\mathcal{R}$. However, the number of mixture components and the corresponding values of the shape parameter vectors also have to be estimated for which we discuss a strategy in the next section. The asymptotic results stated here do not take this form of model selection into account. In Section 5.3 we apply a bootstrap approach to obtain bootstrap confidence intervals for the value of Kendall's τ and Spearman's ρ.

4 Computational details

An efficient multivariate extension of the univariate EM estimation procedure for Erlang mixtures is not straightforward. Indeed, initialization of the parameter values and model selection are the main difficulties when estimating a multivariate Erlang mixture to a data sample and are crucial for its practical use in data analysis. We fill this gap and suggest an effective method to initialize the parameters of a multivariate Erlang mixture and a strategy to select the best set of shape parameter vectors using a model selection criterion.

4.1 Initialization and first run of the EM algorithm

Property 1 ensures that any positive continuous distribution can be approximated by an MME. The formulation of the property also shows how this approximation can be achieved in case the density to be approximated is available. Therefore, it serves as a starting point on how to come up with initial values in case of a sample of observations. A priori, it is however not clear how to translate the property to a finite sample setting.

Initializing data In a finite sample setting, we do not have the underlying density function at our disposal and initialize the parameters making use of an initializing data matrix $\boldsymbol{y}$ of dimension $n \times d$ which contains x_{ij} if the jth element of observation i is uncensored, l_{ij} in the case of right censoring, u_{ij} in the case of left censoring, and $(l_{ij} + u_{ij})/2$ in case of interval censoring. Hence, we use popular simple imputation techniques (see e.g. Leung et al., 1997) to deal with the censoring in the initial step. If the jth element of observation i is missing or right censored at 0, we set y_{ij} equal to missing.

Shapes For any given initial common scale $\theta^{(0)}$, instead of using an infinite set of positive integer shape parameters in each dimension (cfr. Property 1), we restrict this to a maximum number M of shape parameters in each dimension. We select these shape parameters in a sensible way by using M quantiles ranging from the minimum to the maximum in each dimension in order to make a data-driven decision on the locations of the shape parameters. Denoting the p-percent quantile of the initializing data in dimension j by $Q(p; \boldsymbol{y}_j)$, and taking into account that the expected value of a univariate Erlang distribution with shape r and scale θ equals $r\theta$, the set of positive integer shapes in dimension j is chosen as

$$\{r_{1,j}, \ldots, r_{M_j,j}\} = \left\{ \left\lceil \frac{Q(p; \boldsymbol{y}_j)}{\theta^{(0)}} \right\rceil \middle| \, p = 0, \frac{1}{M-1}, \frac{2}{M-1}, \ldots, 1 \right\} . \tag{17}$$

where $\lceil \, \rceil$ denotes upwards rounding, due to the fact that the shapes have to be positive integers. Consequently, several shapes might coincide which results in $M_j \leqslant M$ shape parameters in dimension j. The initial shape set is then constructed as the Cartesian product of the d sets of positive integer shape parameters in each dimension:

$$\mathcal{R} = \{r_{1,1}, \ldots, r_{M_1,1}\} \times \cdots \times \{r_{1,d}, \ldots, r_{M_d,d}\} . \tag{18}$$

Weights The shape parameters in each dimension, multiplied with the common scale parameter $\theta^{(0)}$, form a grid that covers the sample range. As an empirical version of Property 1, the weights $\alpha_{\boldsymbol{r}}$, for each shape parameter vector $\boldsymbol{r} = (r_{m_1,1}, \ldots, r_{m_d,d})$ in $\mathcal{R}$, with $1 \leqslant m_j \leqslant M_j$ for all $j = 1, \ldots, d$, are initialized by the relative frequency of data points in the d-dimensional rectangle $(r_{m_1-1,1}\theta^{(0)}, r_{m_1,1}\theta^{(0)}] \times \cdots \times (r_{m_d-1,d}\theta^{(0)}, r_{m_d,d}\theta^{(0)}]$ defined by the grid:

$$\alpha^{(0)}_{\boldsymbol{r}=(r_{m_1,1},\ldots,r_{m_d,d})} = n^{-1} \sum_{i=1}^{n} \prod_{j=1}^{d} I\left(r_{m_j-1,j}\theta^{(0)} < y_{ij} \leqslant r_{m_j,j}\theta^{(0)}\right), \tag{19}$$

with $r_{0,j} = 0$ for notational convenience and the indicator equal to $1/M_j$ in case y_{ij} is missing. If this hyperrectangle does not contain any data points, the initial weight corresponding to the multivariate Erlang in the mixture with that shape vector will be set equal to zero. Consequently, the weight will remain zero at each subsequent iteration of the EM algorithm (see formulas (12) and (14)). Therefore, these shape vectors can immediately be removed from the set $\mathcal{R}$. At initialization, the truncation is only taken into account to transform the initial values for $\boldsymbol{\alpha}$ into the initial values for $\boldsymbol{\beta}$ via (7).

The maximal number of shape vectors is limited to M^d at the initial step. However, due to the fact that $M_j \leqslant M$ and many shape parameter vectors will receive an initial weight equal to zero, the actual number of shape vectors at the initial step will be lower.

Common scale The initial value of the common scale θ is the most influential for the performance of the initial multivariate Erlang mixture, as is the case in the univariate setting (Verbelen et al., 2015). A value which is too large will result in a multivariate mixture which is too flat (*'underfit'*); a value which is too small will lead to a mixture which is too peaky (*'overfit'*). A priori, it is not evident how one can make an insightful decision on θ. Similar to Verbelen et al. (2015), we therefore introduce an additional tuning parameter: an integer spread factor s. We propose to initialize the common scale as

$$\theta^{(0)} = \frac{\min_j(\max_i(y_{ij}))}{s}. \tag{20}$$

Due to the use of marginal quantiles in (17) , the range of the shape parameters varies according to the sample ranges in each dimension $j = 1, \ldots, d$ with a maximum shape parameter equal to

$$r_{M_j,j} = \left\lceil \frac{\max_i(y_{ij})}{\theta^{(0)}} \right\rceil = \left\lceil \frac{\max_i(y_{ij})}{\min_j(\max_i(y_{ij}))} s \right\rceil . \tag{21}$$

Hence, the spread factor s determines the maximum shape parameter in the dimension with the smallest maximum. The fact that the common scale parameter is equal across all dimensions is compensated by the different choice of the shape parameters in each dimension based on marginal quantiles. This ensures that the initialization works well when the ranges in each dimension are different and also gives reasonable initial approximations in case the data are skewed.

Algorithm 1 EM algorithm for a multivariate Erlang mixture.

{**Initial step**}
Choose M and s
θ as in (20)
Compute: shape parameters in each dimension as in (17) and shape set $\mathcal{R}$ as in (18)
mixture weights $\boldsymbol{\alpha}$ as in (19)
$\mathcal{R} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\alpha_{\boldsymbol{r}} \neq 0\}$
Transform weights $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$ as in (7)
{**EM algorithm**}
while log-likelihood (6) improves **do**
{**E-step**}
Compute: posterior probabilities (12)
conditional expectations (13)
{**M-step**}
Update: weights $\boldsymbol{\beta}$ as in (14)
scale θ by numerically solving (15)
end while
Transform weights $\boldsymbol{\beta}$ to $\boldsymbol{\alpha}$ using (16)
return $\text{MME}_{init} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$

Apply EM algorithm Given an initial choice for the set $\mathcal{R}$ of shape parameter vectors, the initial common scale estimate $\theta^{(0)}$ and the initial weights $\boldsymbol{\beta}^{(0)} = \{\beta_{\boldsymbol{r}}^{(0)} \,|\boldsymbol{r} \in \mathcal{R}\}$, we find the maximum likelihood estimators for $(\boldsymbol{\beta}, \theta)$ corresponding to this initial multivariate mixtures of Erlangs, denoted by MME_{init}, via the EM algorithm as explained in section 3.3. An overview of the initialization and the EM algorithm written in pseudo code is given in Algorithm 1.

4.2 Reduction of the shape vectors

The initial shape set $\mathcal{R}$ might not be optimal. After application of the EM algorithm, we reduce the number of mixture components of the fitted multivariate Erlang mixture. We use a backward stepwise search based on an information criterion. Information criteria, such as Akaike's information criterion (AIC, Akaike, 1974) and Schwartz's Bayesian information criterion (BIC, Schwarz, 1978), measure the quality of the model as a trade-off between the goodness-of-fit, via the log-likelihood, and the model complexity, via the number of parameters in the model. Models with a smaller value of the information criterion are preferred. Based on numerical experiments, we prefer the use of BIC over AIC since it has a stronger penalty term for the number of parameters in the model and hence leads to more parsimonious models. BIC is computed as

$$\text{BIC} = -2 \cdot l(\boldsymbol{\Theta}; \mathcal{X}) + \ln(n) \cdot |\mathcal{R}| \cdot (d+1)\,, \tag{22}$$

where $|\mathcal{R}|$ indicates the number of shape parameter vectors in the shape set $\mathcal{R}$.

We reduce the number of mixture components by removing all redundant shape vectors from the initial mixture based on BIC. In the backward selection strategy, depicted in pseudo code in Algorithm 2, we delete the shape parameter vector $\boldsymbol{r}$ from the set $\mathcal{R}$ for which the corresponding mixture component has the smallest weight $\beta_{\boldsymbol{r}}$. The remaining weights are standardized to sum to one. Along with the previous maximum likelihood estimate for the common scale, they serve as initial estimates to find the maximum likelihood estimators for $(\boldsymbol{\beta}, \theta)$ corresponding to the reduced set $\mathcal{R}_{red}$ of shape parameter vectors by again applying the EM algorithm. In case this maximum likelihood estimate achieves a lower BIC value, the reduced set $\mathcal{R}_{red}$ of shape parameters is accepted and we reduce the number of components further in the same manner. If not, we keep the previous set. This backward approach provides efficient initial parameter estimates for the reduced set of shape parameter vectors and ensures a fast convergence of the EM algorithm.

4.3 Adjustment of the shape vectors

In a next step we improve the shape parameter vectors of the remaining Erlang components in the mixture. Each time we adjust one of the components of a shape parameter vector by shifting its value by one (increase or decrease) and use the maximum likelihood estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\theta})$ corresponding to the current shape parameter set $\mathcal{R}$ as initial values $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{adj}$ of the mixture

Algorithm 2 Reduction of the shape vectors

input $\text{MME}_{init} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$
while BIC (22) improves **and** $|\mathcal{R}| > 1$ **do**
  $\mathcal{R}_{red} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\beta_{\boldsymbol{r}} \neq \min_{\boldsymbol{r}\in\mathcal{R}} \beta_{\boldsymbol{r}}\}$
  $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{red} \leftarrow (\{\beta_{\boldsymbol{r}} / \sum_{\boldsymbol{r}\in\mathcal{R}_{red}} \beta_{\boldsymbol{r}} \,|\boldsymbol{r} \in \mathcal{R}_{red}\}, \theta)$
  Compute MLE for $(\boldsymbol{\beta}, \theta)_{red}$ using the EM algorithm with initial values $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{red}$
  if BIC (22) improves **then**
    $\mathcal{R} \leftarrow \mathcal{R}_{red}$
    $(\boldsymbol{\beta}, \theta) \leftarrow (\boldsymbol{\beta}, \theta)_{red}$
  end if
end while
return $\text{MME}_{red} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$

of Erlang distributions with slightly adjusted shape parameter vector set $\mathcal{R}_{adj}$. These initial values are close to the maximum likelihood estimates which guarantees fast convergence. In case the maximum likelihood estimate corresponding to the adjusted set $\mathcal{R}_{adj}$ achieves a lower log-likelihood value (6), the adjusted set $\mathcal{R}_{adj}$ is accepted and we continue adjusting the value of the shape parameter in the same direction. If not, we keep the previous set of shape parameter combinations.

The gradual adjustment strategy of the shape parameter combinations is described in detail in Algorithm 3. While the log-likelihood improves, we continue to consecutively increase or decrease the value of a component of a shape parameter vector if it leads to a better fit. The algorithm converges when no single addition or subtraction of the value of any of the components of any of the shape parameter vectors leads to an improvement in the log-likelihood.

After adjusting the shape parameters, we apply the reduction step in combination with the adjustment step. Based on BIC we further reduce the number of shape parameter vectors by deleting the shape vector with the smallest mixture weight and adjusting the values of the remaining ones. The outline of this adjustment and further reduction of the shape parameter vectors, which results in the final MME, is given in Algorithm 4.

Algorithm 3 Adjustment of the shape combinations

input $\text{MME}_{red} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$
while log-likelihood (6) improves **do**
  for $j \in \{1, \ldots, d\}$ **do**
    for $\widetilde{\boldsymbol{r}} \in \mathcal{R}$ **do**
      repeat
        if $(\widetilde{r}_1, \ldots, \widetilde{r}_j + 1, \ldots, \widetilde{r}_d) \notin \mathcal{R}$ **then**
          $\mathcal{R}_{adj} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\boldsymbol{r} \neq \widetilde{\boldsymbol{r}}\} \cup \{(\widetilde{r}_1, \ldots, \widetilde{r}_j + 1, \ldots, \widetilde{r}_d)\}$
          Compute MLE for $(\boldsymbol{\beta}, \theta)_{adj}$ using the EM algorithm with initial values $(\boldsymbol{\beta}, \theta)$
          if log-likelihood (6) improves **then**
            $\mathcal{R} \leftarrow \mathcal{R}_{adj}$
            $(\boldsymbol{\beta}, \theta) \leftarrow (\boldsymbol{\beta}, \theta)_{adj}$
          end if
        end if
      until $(\widetilde{r}_1, \ldots, \widetilde{r}_j + 1, \ldots, \widetilde{r}_d) \in \mathcal{R}$ **or** log-likelihood (6) no longer improves
    end for
    for $\widetilde{\boldsymbol{r}} \in \mathcal{R}$ **do**
      repeat
        if $(\widetilde{r}_1, \ldots, \widetilde{r}_j - 1, \ldots, \widetilde{r}_d) \notin \mathcal{R}$ **and** $\widetilde{r}_j - 1 \geqslant 1$ **then**
          $\mathcal{R}_{adj} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\boldsymbol{r} \neq \widetilde{\boldsymbol{r}}\} \cup \{(\widetilde{r}_1, \ldots, \widetilde{r}_j - 1, \ldots, \widetilde{r}_d)\}$
          Compute MLE for $(\boldsymbol{\beta}, \theta)_{adj}$ using the EM algorithm with initial values $(\boldsymbol{\beta}, \theta)$
          if log-likelihood (6) improves **then**
            $\mathcal{R} \leftarrow \mathcal{R}_{adj}$
            $(\boldsymbol{\beta}, \theta) \leftarrow (\boldsymbol{\beta}, \theta)_{adj}$
          end if
        end if
      until $(\widetilde{r}_1, \ldots, \widetilde{r}_j - 1, \ldots, \widetilde{r}_d) \in \mathcal{R}$ **or** $\widetilde{r}_j - 1 = 0$ **or** log-likelihood (6) no longer improves
    end for
  end for
end while
return $\text{MME}_{adj} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$

Algorithm 4 Adjustment and further reduction of the shape vectors

input $\text{MME}_{adj} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$
while BIC (22) improves **and** $|\mathcal{R}| > 1$ **do**
  $\mathcal{R}_{red} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\beta_{\boldsymbol{r}} \neq \min_{\boldsymbol{r} \in \mathcal{R}} \beta_{\boldsymbol{r}}\}$
  $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{red} \leftarrow (\{\beta_{\boldsymbol{r}} / \sum_{\boldsymbol{r} \in \mathcal{R}_{red}} \beta_{\boldsymbol{r}} \,|\boldsymbol{r} \in \mathcal{R}_{red}\}, \theta)$
  Compute MLE for $(\boldsymbol{\beta}, \theta)_{red}$ using the EM algorithm with initial values $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{red}$
  Apply adjustment algorithm 3
  if BIC (22) improves **then**
    $\mathcal{R} \leftarrow \mathcal{R}_{adj}$
    $(\boldsymbol{\beta}, \theta) \leftarrow (\boldsymbol{\beta}, \theta)_{adj}$
  end if
end while
return $\text{MME}_{adj} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$

5 Examples

We demonstrate the proposed fitting procedure on three datasets, each time highlighting a different aspect of multivariate mixtures of Erlangs. In a first simulated two-dimensional example, we explicitly illustrate the different steps of the estimation procedure. Second, we model the waiting time between eruptions and the duration of the eruptions of the old faithful geyser dataset. Based on the fitted two-dimensional MME, we immediately obtain the distribution of the sum of the waiting time and the duration, representing the total cycle time. In the third example, we use multivariate mixtures of Erlangs to model the udder infection times of dairy cows observed in a mastitis study, and use the fitted MME to analytically quantify the positive correlation between the udder infection times using the explicit expression of the bivariate measures of association Kendall's tau and Spearman's rho in the MME setting.

The resulting MME after applying the different steps in choosing the shape vectors depends heavily on the starting values. Therefore it is crucial to sufficiently explore the effect of changing the value of the tuning parameters M and s and compare the results of several different initial starting points for the shape set. In addition to the value of BIC, graphs aid the assessment of the fitted model.

5.1 Simulated data

As a first example, we generate 1000 uncensored and untruncated observations from a bivariate normal copula with correlation coefficient 0.75 and Erlang distributed margins with shape parameter equal to 2 and 10, respectively, and scale parameter equal to 3 and 20, resp. A scatterplot of this simulated dataset is shown in Figure 1a. Due to the parameter choice, the ranges in each dimension are quite different.

We now apply the different steps of the estimation procedure on this dataset and graphically illustrate the interpretations and effects of these steps. First we consider the initialization strategy for the shape set $\mathcal{R}$, the scale parameter θ and the mixture weights $\boldsymbol{\beta}$, based on the denseness property of MME in Property 1, as explained in Section 4.1. This strategy is controlled by two tuning parameters, a maximum number M of shape parameters in each dimension and a spread factor s. In this illustration, we use $M = 10$ and $s = 20$. For this choice, the scale θ is initialized as

$$\theta^{(0)} = \frac{\min_j(\max_i(x_{ij}))}{s} = \frac{27.32452}{20} = 1.366226\,.$$

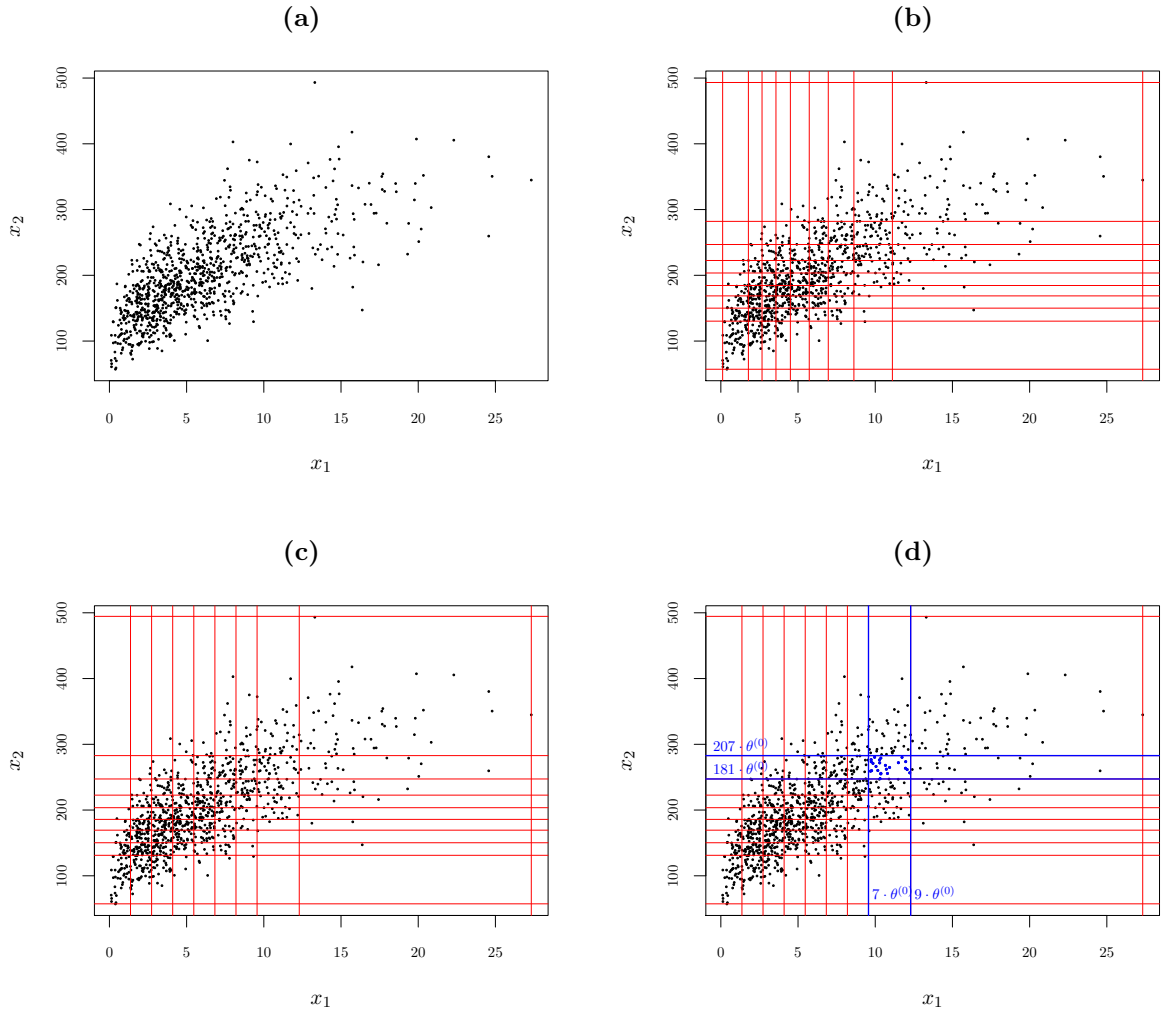

Figure 1: Simulated example: (a) scatterplot, (b) marginal quantile grid, (c) grid formed by multiplying the shapes (17) by the common scale (20) and (d) initial weight $\alpha^{(0)}_{\boldsymbol{r}=(9,207)} = 0.024$.

In order to make a data driven choice for the initial positions of the shape parameters, we compute M marginal quantiles in each dimension, which are depicted in Figure 1b and form a grid that covers the data range. These marginal quantiles are then divided by the initial scale $\theta^{(0)}$ and rounded upwards to initialize the shape parameters in each dimension:

$$\{r_{1,j}, \ldots, r_{M_j,j}\} = \left\{ \left\lceil \frac{Q(p; \boldsymbol{x}_j)}{\theta^{(0)}} \right\rceil \middle| p = 0, \frac{1}{9}, \frac{2}{9}, \ldots, 1 \right\} \quad \text{for } j = 1, 2\,.$$

The shape set $\mathcal{R}$ is constructed as the Cartesian product of the set of shape parameters in each

dimension:

$$\begin{aligned}\mathcal{R} &= \{r_{1,1}, \ldots, r_{M_1,1}\} \times \{r_{1,2}, \ldots, r_{M_2,2}\} \\ &= \{1, 2, 3, 4, 5, 6, 7, 9, 20\} \times \{42, 96, 110, 124, 136, 149, 163, 181, 207, 362\}\,.\end{aligned}$$

Due to the rounding, shape 2 appears twice in the first dimension and only 9 instead of 10 shapes remain in that dimension. Due to the choice of $\theta^{(0)}$, $s = 20$ is the maximal shape parameter in the first dimension, the dimension with the smallest maximum. The maximal shape in the second dimension is s times the ratio of the maximum in the second dimension and the lowest maximum, rounded upwards (see (21)). If we multiply this shape set $\mathcal{R}$ with the initial scale $\theta^{(0)}$, we obtain a grid that covers the entire sample range which is depicted in Figure 1c. This grid differs from the marginal quantile grid due to the rounding and is used to initialize the weights as the relative frequency of data points in the 2-dimensional rectangle corresponding to each shape vector:

$$\alpha^{(0)}_{\boldsymbol{r}=(r_{m_1,1}, r_{m_2,2})} = 0.001 \sum_{i=1}^{1000} \prod_{j=1}^{2} I\left(r_{m_j-1,j}\theta^{(0)} < y_{ij} \leqslant r_{m_j,j}\theta^{(0)}\right)\,.$$

For example, for the shape vector $\boldsymbol{r} = (r_{m_1,1}, r_{m_2,2}) = (9, 207)$, we consider the 2-dimensional rectangle $(r_{m_1-1,1}\theta^{(0)}, r_{m_1,1}\theta^{(0)}] \times (r_{m_2-1,2}\theta^{(0)}, r_{m_2,2}\theta^{(0)}] = (7 \cdot \theta^{(0)}, 9 \cdot \theta^{(0)}] \times (181 \cdot \theta^{(0)}, 207 \cdot \theta^{(0)}]$ shown in Figure 1d, leading to an initial weight of

$$\alpha^{(0)}_{\boldsymbol{r}=(9,207)} = 0.001 \sum_{i=1}^{1000} I\left(7 \cdot \theta^{(0)} < y_{i1} \leqslant 9 \cdot \theta^{(0)}\right) I\left(181 \cdot \theta^{(0)} < y_{i2} \leqslant 207 \cdot \theta^{(0)}\right) = 0.024\,,$$

since 24 of the 1000 observations lie in this rectangle. The resulting initial MME contains 71 shape vectors with a nonzero weight and already forms a reasonable approximation for the main portion of the data. In Figure 2a, we show the scatterplot of the data with an overlay of the density of the initial MME using a contour plot and heat map. In the margins, we plot the marginal histograms with an overlay of the true densities in blue and the fitted densities in red. In the second dimension, there is too much weight in the tail and too little near the origin. After applying the EM algorithm a first time with these initial estimates, we obtain the maximum likelihood estimates of the weights and scale corresponding to this choice of the shape set (Section 4). In Figure 2b, we observe that the fit is better in the tail, but there is still too

little weight in the second dimension near the origin, due to a bad positioning of the first shape in second dimension.

Table 1: Parameter estimates of the MME with 11 mixture components fitted to the simulated data.

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(1, 56)	0.0124	1.2889
(2, 84)	0.0814	
(3, 112)	0.1773	
(3, 132)	0.1005	
(4, 143)	0.1568	
(4, 164)	0.0257	
(5, 164)	0.1320	
(6, 189)	0.1586	
(8, 223)	0.1097	
(11, 273)	0.0446	
(11, 382)	0.0010	

Hence, the initial set of shape parameter vectors is not ideal and additional steps are required to improve the shape set. First, we reduce the number of mixture components from 71 to 17 by subsequently removing the mixture component having the smallest weight if it is found to be redundant based on BIC (Section 4.2). The fit of this reduced mixture in Figure 2c nearly coincides with the one in Figure 2b. Second, we adjust the values of the shape parameter vectors and further reduce the number of mixture components based on BIC (Section 4.3) until we obtain a close-fitting MME with 11 shape parameter vectors (Figure 2d). The parameter estimates of this final MME are given in Table 1.

5.2 Old faithful geyser data

We consider the waiting time between eruptions and the duration of the eruption for the Old faithful geyser in Yellowstone National Park, Wyoming, USA. We use the version of Azzalini and Bowman (1990) which contains 299 observations. This dataset is popular in the field of nonparametric density estimation (see e.g. Silverman, 1986; Härdle, 1991). We stress that we use MME as a multivariate density estimation technique, and not as a mixture modeling technique to identify subgroups in this data.

We fit a two-dimensional MME to the data using the fitting strategy explained in Section 4. We perform a grid search to identify good values for the tuning parameters M and s. We let s vary between 10 and 90 by 10 and between 100 and 1000 by 100 and set M equal to 5, 10 and 20.

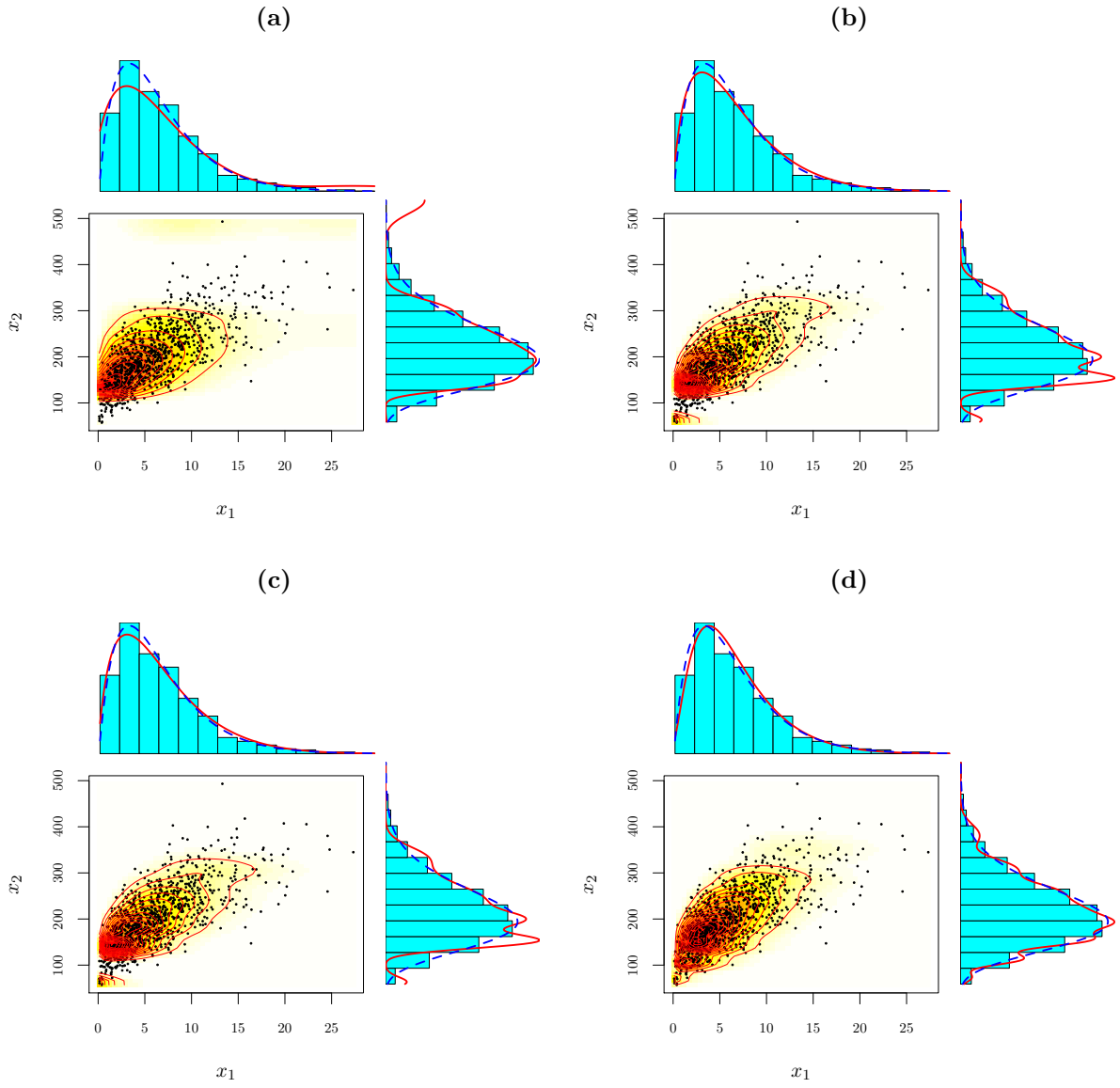

Figure 2: Scatterplot of the simulated data with an overlay of the fitted density of the MME using a contour plot and heat map. In the margins, we plot the marginal histograms with an overlay of the true densities in blue and the fitted densities in red. In (a), we display the fit after initialization, in (b) after applying the EM algorithm a first time, in (c) after applying the reduction step and in (d) after applying the adjustment and further reduction step.

To illustrate the importance and effect of the tuning parameters, we report part of the results of the search grid, up to $s = 200$, in Figure 3 and Table 2. Values of s beyond 200 resulted in MME which were overfitting the data.

Table 2: BIC values and number of mixture components when fitting an MME to the Old Faithful geyser data, starting from different values of the tuning parameters. The minimum BIC value is underlined and obtained for $M = 10$ and $s = 90$.

	$M = 5$		$M = 10$		$M = 20$	
s	BIC	$\lvert\mathcal{R}\rvert$	BIC	$\lvert\mathcal{R}\rvert$	BIC	$\lvert\mathcal{R}\rvert$
10	3211.134	2	3211.134	2	3211.134	2
20	3133.564	5	3148.824	5	3148.824	5
30	3069.731	6	3069.731	6	3083.757	6
40	3056.588	8	3024.869	9	3051.427	6
50	3026.997	8	3011.941	12	3023.951	15
60	3011.567	8	3008.350	14	3040.962	16
70	3008.319	8	3008.350	14	3018.867	15
80	3015.743	8	3007.694	15	3039.017	17
90	3028.742	8	<u>2998.870</u>	15	3047.314	18
100	3029.431	8	3005.343	15	3023.761	17
200	3037.532	8	3026.490	23	3224.578	36

The resulting MME depends on the value of the tuning parameters. However, multiple MME can result in a satisfactory fit of the data. BIC indicates that the best-fitting MME is obtained for $M = 10$ and $s = 90$. The parameter estimates of this MME are reported in Table 3. Both the marginals as well as the dependence structure are adequately represented by this MME as is confirmed graphically in Figure 4a. Since the maximum of the waiting times is about 20 times as big as the maximum of the duration times whereas the scale parameter of the MME is the same across dimensions, the fitted marginal density is more capricious in the dimension of the waiting times and smoother in the dimension of the duration times.

We are interested in the distribution of the duration of the total cycle, i.e. the sum of the waiting time until the eruption and the duration of the eruption. Based on the fitted two-dimensional MME and due to the analytical properties of MME, we immediately obtain the distribution of this sum, which is a univariate mixture of Erlang distributions with the same scale, the sum of the shape parameters across the dimensions as shape parameters and the same corresponding weights in (Lee and Lin, 2012, Theorem 5.1). Hence, the parameters of this univariate mixture of Erlang distributions are readily available from Table 3. Comparing the histogram of the observed total times to the fitted density in Figure 4b reveals a close approximation.

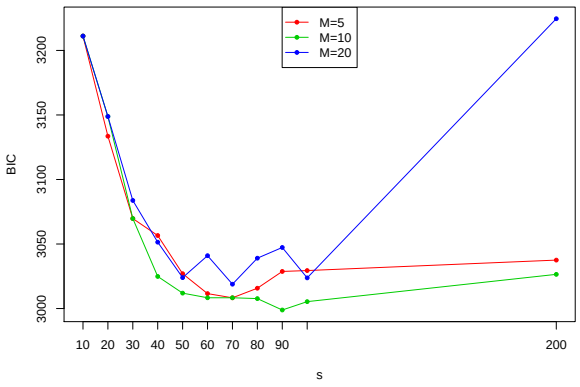

Figure 3: BIC values when fitting an MME to the Old Faithful geyser data, starting from different values of the tuning parameters. The minimum BIC value is obtained for $M = 10$ and $s = 90$.

Table 3: Parameter estimates of the best-fitting MME with 15 mixture components fitted to the Old Faithful geyser data.

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(791, 79)	0.0061	0.0556
(893, 81)	0.1103	
(964, 79)	0.0798	
(1047, 77)	0.0795	
(1121, 83)	0.0378	
(1193, 79)	0.0402	
(1314, 74)	0.0893	
(1319, 37)	0.0387	
(1418, 73)	0.1284	
(1425, 36)	0.1380	
(1543, 73)	0.0633	
(1551, 36)	0.1249	
(1660, 72)	0.0142	
(1672, 34)	0.0462	
(1940, 36)	0.0033	

(a) **(b)**

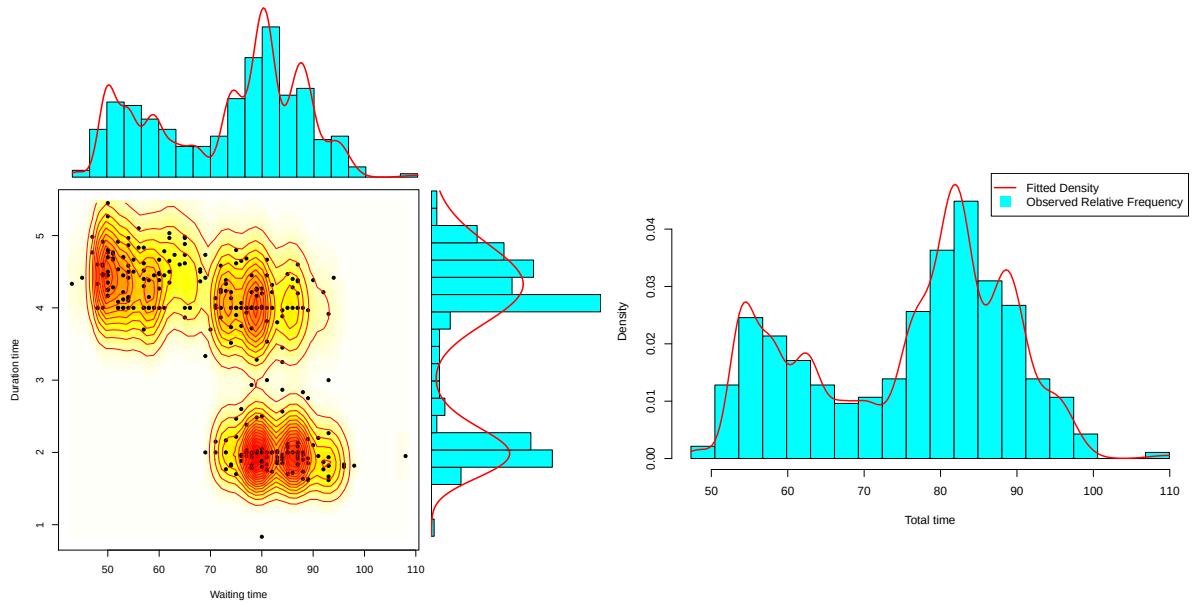

Figure 4: Graphical evaluation of the best-fitting MME to the Old Faithful geyser data. In (a), we display the scatterplot of the data with an overlay of the fitted density using a contour plot and heat map. The margins show the marginal histograms with an overlay of the fitted densities in red. In (b), we compare the fitted density of the sum of the components and the histogram of the observed total cycle times.

5.3 Mastitis study

Mastitis is economically one of the most important diseases in the dairy sector since it leads to reduced milk yield and milk quality. In this example, we consider infectious disease data from a mastitis study by Laevens et al. (1997). This dataset has also been used in Goethals et al. (2009) and Ampe et al. (2012).

We focus on the infection times of individual cow udder quarters with a bacterium. As each udder quarter is separated from the three other quarters, one quarter might be infected while the other quarters remain infection-free. However, the dependence must be modeled since the data are hierarchical, with individual observations at the udder quarter level being correlated within the cow. Additionally, the infection times are not known exactly due to a period follow-up, which is often the case in observational studies since a daily checkup would not be feasible. Roughly each month, the udder quarters are sampled and the infection status is assessed, from the time of parturition, at which the cow was included in the cohort and assumed to be infection-free, until the end of the lactation period. This generates interval-censored data since for udder quarters that experience an event it is only known that the udder quarter got infected between the last visit at which it was infection-free and the first visit at which it was infected. Observations can also be right censored if no infection occurred before the end of the lactation period, which is roughly 300-350 days but different for every cow, before the end of the study or if the cow is lost to follow-up during the study, for example due to culling.

The data we consider contains information on 100 dairy cows on the time to infection of the four udder quarters by different types of bacteria. This dataset is used in Goethals et al. (2009), who model the data using an extended shared gamma frailty model that is able to handle the interval censoring and clustering simultaneously. We treat the infection times at the udder quarter level of the cow as four-dimensional interval and right censored data of which we estimate the underlying density using MME. The udder quarters are denoted as RL (rear left), FL (front left), RR (rear right) and FR (front right).

In search for the best values of the tuning parameters in the MME estimation procedure, we first fixed $M = 20$ and let s vary between 10 and 100 by 10 and between 100 and 1000 by 100. As the best final fit was obtained for $s = 10$, we varied M between 10 and 100 by 10 for s fixed at 10. The resulting fits did, however, not depend on M when s is as low as 10 since the starting values were identical. Varying s from 5 tot 15 for $M = 20$ confirmed that the best fit is obtained

for $M = 20$ and $s = 10$. For this setting, the initial number of shape vectors was 73, which got reduced to 6 after the reduction step and to 4 after the adjustment step. The final parameter estimates of the best-fitting mixture are given in Table 4.

Table 4: Parameter estimates of the best-fitting MME with four mixture components fitted to the mastitis data (infections by all bacteria).

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(2, 2, 2, 2)	0.4897	37.8621
(3, 5, 8, 4)	0.1331	
(7, 5, 2, 7)	0.2262	
(10, 14, 11, 8)	0.1510	

In order to graphically examine the goodness-of-fit of the fitted MME, we construct in Figure 5 a generalization of the scatterplot matrix. On the diagonal we compare the Turnbull non-parametric estimate of the survival curve for right and interval censored data (Turnbull, 1976), along with the log-transformed equal precision simultaneous confidence intervals (Nair, 1984), to the univariate marginal survival function of the fitted MME. On the off-diagonal, we construct bivariate scatterplots of interval and right censored data points, represented using the effective visualization of Li et al. (2015). Interval censored observation are depicted as segments or rectangles ranging from the lower to the upper censoring points and right censored observations are depicted as arrows starting from the lower censoring point and pointing to the censoring direction. On top, we display the contour plot and heat map representing the density of the bivarite marginal of the fitted MME. Based on this graph, we observe that in four dimensions, with 100 interval and right censored observations, we are able to fit an MME with four shape parameter vectors which appropriately captures the marginals as well as the dependence structure.

As a measure of the infectivity of the agent causing the disease, we are interested in the correlation between udder infection times. Due to the fact that the bivariate marginals again belong to the MME class and the analytical qualities of MME, we have closed-form expressions for Kendall's τ and Spearman's ρ (Lee and Lin, 2012, Theorem 3.2 and 3.3). Note that these do not depend on the common scale parameter. For the interval and right censored sample, we can hence estimate these measures based on the fitted MME to analytically quantify the positive correlation between each pair of udder quarter infection times (Table 5).

Inference is not straightforward due to the model selection as pointed out in Section 3.3. In order to quantify the uncertainty and construct an approximate confidence interval for the bivariate

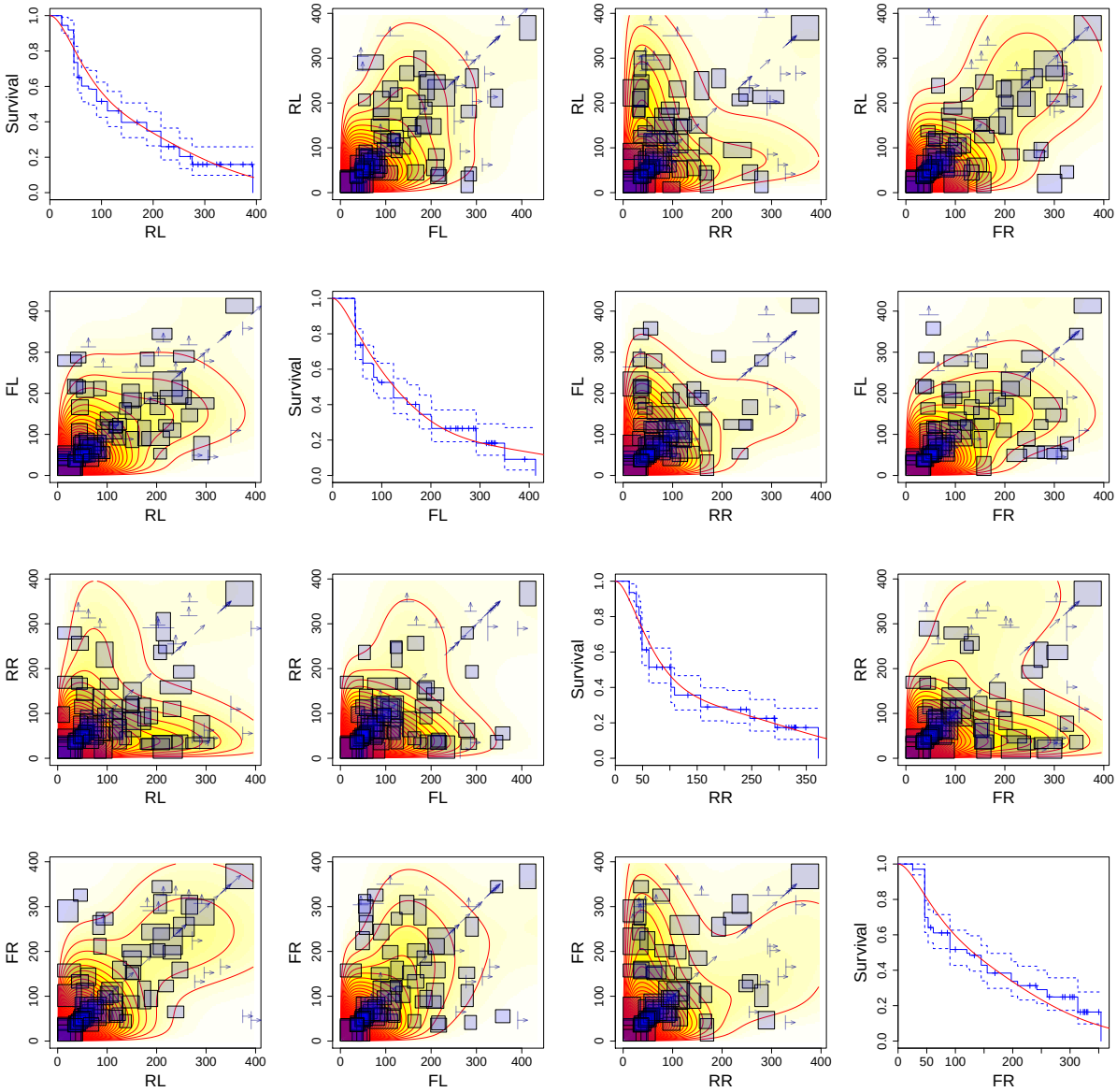

Figure 5: Scatterplot matrix comparing the fitted four-dimensional MME to the observed interval and right censored observations of the mastitis data (infections by all bacteria). For more explanation, see Section 5.3

measures of association, we resort to a bootstrapping procedure (Efron and Tibshirani, 1994). By sampling with replacement from the original four-dimensional dataset of size 100, we generate 1000 bootstrap samples of the same size 100. For each of these bootstrap samples, we fit an MME where we set the tuning parameter M equal to 20 and let s vary between 5 and 25. We choose this fixed grid for each bootstrap sample since the optimal tuning parameters for the full sample were $M = 20$ and $s = 10$ and the starting values are not that sensitive with respect

to M for low values of s. We thereby obtain 1000 estimates for each measure of association. The 5% and 95% quantiles of these estimates are used to construct a 90% bootstrap percentile confidence interval for each Kendall's τ and Spearman's ρ in Table 5.

Table 5: Estimates and 90% bootstrap confidence intervals for the bivariate measures of association Kendall's τ and Spearman's ρ based on the fitted MME for the mastitis data (infections by all bacteria).

		RL	FL	RR
FL	τ	0.4187		
		$(0.3329, 0.5515)$		
	ρ	0.6019		
		$(0.4727, 0.7439)$		
RR	τ	0.2018	0.3307	
		$(0.1693, 0.3989)$	$(0.2585, 0.4784)$	
	ρ	0.3004	0.4852	
		$(0.2423, 0.5616)$	$(0.3806, 0.6664)$	
FR	τ	0.4326	0.4105	0.2119
		$(0.3598, 0.5538)$	$(0.2701, 0.4883)$	$(0.1543, 0.3968)$
	ρ	0.6354	0.5994	0.3122
		$(0.5066, 0.7608)$	$(0.3875, 0.6794)$	$(0.2206, 0.5577)$

6 Discussion

MME form a highly flexible class of distributions which are at the same time mathematically tractable. From Property 1, we know that any positive continuous multivariate distribution can be approximated up to any accuracy by an infinite multivariate mixture of Erlang distributions. Our contribution presents a computationally efficient initialization and adjustment strategy for the shape parameter vectors, translating this theoretical aspect in a strong point in practice as well. In the examples, we demonstrate how the fitting procedure is able to estimate an MME that adequately represents both the marginals and the dependence structure. By extending the EM algorithm, we are now also able to deal with left, interval or right censored and truncated data. MME therefore form a valuable multivariate density estimation technique to analyze realistic data, even in incomplete data settings, and to model the dependence directly in a low dimensional setting.

Their tractability allows to derive explicit expression of properties of interest. Willmot and Woo (2015) have paved the way for applying MME in insurance loss modeling, survival analysis and ruin theory. When modeling insurance losses or dependent risks from different portfolios

or lines of buniness using MME, the aggregate and excess losses have again a univariate and multivariate mixture of Erlangs distribution. Stop-loss moments, several types of premiums, risk capital allocation based on the Tail-Value-at-Risk (TVaR) or covariance rule for regulatory risk capital requirements (see e.g. Dhaene et al., 2012) have analytical expressions. When modeling bivariate lifetimes and pricing joint-life and last-survivor insurance (see e.g. Frees et al., 1996) using MME, the distribution of the minimum and maximum is again a univariate mixture of Erlangs. Such kind of data are always left truncated and right censored. The extension of the fitting procedure for MME presented in this paper, allows to take the right censoring into account. Left truncation can only be properly handled when the left truncation points are fixed for each observation. This is however not the case when pricing joint-life and last-survivor insurance since the ages at which policyholders enter a contract vary.

The reduction and adjustment steps of the shape parameters in the fitting procedure iteratively make use of the EM algorithm and can be time consuming. Further adjustment is needed to estimate parameters in high dimensional settings. As also acknowledged in the univariate case (Verbelen et al., 2015), the modeling of heavy-tailed distributions using MME is challenging since MME are not able to extrapolate the heaviness in the tail.

Acknowledgements

The authors wish to thank Dr. H. Laevens (Catholic University College Sint-Lieven, Sint-Niklaas, Belgium), for permission to use the mastitis data, and the referees for their comments. This work was supported by the agency for Innovation by Science and Technology IWT, IAP Research Network P6/03 of the Belgian State (Belgian Research Policy) and by KU Leuven grant GOA/12/14.

Appendix A Partial derivative of Q

In order to maximize $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ with respect to θ, we set the first order partial derivative at $\theta^{(k)}$ equal to zero. In the second equation, expression (2) of the cumulative distribution of an Erlang, while (14) is used to obtain the third equation.

$$
\begin{aligned}
&\frac{\partial Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)})}{\partial\theta}\bigg|_{\theta=\theta^{(k)}} \\
&\quad= \sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \left(\frac{\sum_{j=1}^{d} E\left(X_{ij}\,\middle|\,Z_{i\boldsymbol{r}}=1,l_{ij},u_{ij},t_j^l,t_j^u;\theta^{(k-1)}\right)}{\theta^2} - \frac{\sum_{j=1}^{d} r_j}{\theta} \right. \\
&\qquad \left. - \sum_{j=1}^{d} \frac{\frac{\partial}{\partial\theta}\left[F(t_j^u;r_j,\theta)-F(t_j^l;r_j,\theta)\right]}{F(t_j^u;r_j,\theta)-F(t_j^l;r_j,\theta)} \right)\bigg|_{\theta=\theta^{(k)}} \\
&\quad= \frac{1}{\theta^2}\sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} E\left(X_{ij}\,\middle|\,Z_{i\boldsymbol{r}}=1,l_{ij},u_{ij},t_j^l,t_j^u;\theta^{(k-1)}\right) - \frac{n}{\theta}\sum_{\boldsymbol{r}\in\mathcal{R}}\left(\frac{\sum_{i=1}^{n} z_{i\boldsymbol{r}}^{(k)}}{n}\right)\sum_{j=1}^{d} r_j \\
&\qquad - \sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} \frac{\frac{\partial}{\partial\theta}\left(\gamma(r_j,t_j^u/\theta)-\gamma(r_j,t_j^l/\theta)\right)}{(r_j-1)!\left(F(t_j^u;r_j,\theta)-F(t_j^l;r_j,\theta)\right)}\bigg|_{\theta=\theta^{(k)}} \\
&\quad= \frac{1}{\theta^2}\sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} E\left(X_{ij}\,\middle|\,Z_{i\boldsymbol{r}}=1,l_{ij},u_{ij},t_j^l,t_j^u;\theta^{(k-1)}\right) - \frac{n}{\theta}\sum_{\boldsymbol{r}\in\mathcal{R}}\beta_{\boldsymbol{r}}^{(k)}\sum_{j=1}^{d} r_j \\
&\qquad - \frac{n}{\theta^2}\sum_{\boldsymbol{r}\in\mathcal{R}}\beta_{\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} \frac{\left(t_j^l\right)^{r_j} e^{-t_j^l/\theta} - \left(t_j^u\right)^{r_j} e^{-t_j^u/\theta}}{\theta^{r_j-1}(r_j-1)!\left(F(t_j^u;r_j,\theta)-F(t_j^l;r_j,\theta)\right)}\bigg|_{\theta=\theta^{(k)}} = 0\,,
\end{aligned}
$$

where we used expression (2) of the cumulative distribution of an Erlang in the second equality and (14) in the third.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Ampe, B., Goethals, K., Laevens, H., and Duchateau, L. (2012). Investigating clustering in interval-censored udder quarter infection times in dairy cows using a gamma frailty model. *Preventive veterinary medicine*, 106(3):251–257.

Asmussen, S., Nerman, O., and Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, pages 419–441.

Assaf, D., Langberg, N. A., Savits, T. H., and Shaked, M. (1984). Multivariate phase-type distributions. *Operations Research*, 32(3):688–702.

Azzalini, A. and Bowman, A. (1990). A look at some data on the old faithful geyser. *Applied Statistics*, pages 357–365.

Cossette, H., Côté, M.-P., Marceau, E., and Moutanabbir, K. (2013a). Multivariate distribution defined with Farlie–Gumbel–Morgenstern copula and mixed Erlang marginals: Aggregation and capital allocation. *Insurance: Mathematics and Economics*, 52(3):560–572.

Cossette, H., Mailhot, M., Marceau, E., and Mesfioui, M. (2013b). Bivariate lower and upper orthant Value-at-Risk. *European Actuarial Journal*, 3(2):321–357.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dhaene, J., Tsanakas, A., Valdez, E. A., and Vanduffel, S. (2012). Optimal capital allocation principles. *Journal of Risk and Insurance*, 79(1):1–28.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Eisele, K.-T. (2005). EM algorithm for bivariate phase distributions. In *ASTIN Colloquium, Zurich, Switzerland. http://www. actuaries. org/ASTIN/Colloquia/Zurich/Eisele. pdf*.

Frees, E. W., Carriere, J., and Valdez, E. (1996). Annuity valuation with dependent mortality. *Journal of Risk and Insurance*, pages 229–261.

Georges, P., Lamy, A.-G., Nicolas, E., Quibel, G., and Roncalli, T. (2001). Multivariate survival modelling: A unified approach with copulas. *Unpublished paper, Groupe de Recherche Operationnelle, Credit Lyonnais, France*.

Goethals, K., Ampe, B., Berkvens, D., Laevens, H., Janssen, P., and Duchateau, L. (2009). Modeling interval-censored, clustered cow udder quarter infection times through the shared gamma frailty model. *Journal of agricultural, biological, and environmental statistics*, 14(1):1–14.

Härdle, W. (1991). *Smoothing techniques: with implementation in S*. Springer.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*, volume 73. CRC Press.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2013). *Loss models: Further topics*. John Wiley & Sons.

Laevens, H., Deluyker, H., Schukken, Y., De Meulemeester, L., Vandermeersch, R., De Muelenaere, E., and De Kruif, A. (1997). Influence of parity and stage of lactation on the somatic cell count in bacteriologically negative dairy cows. *Journal of Dairy Science*, 80(12):3219–3226.

Lee, G. and Scott, C. (2012). EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816 – 2829.

Lee, S. and McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24(2):181–202.

Lee, S. C. and Lin, X. S. (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal*, 14(1):107–130.

Lee, S. C. and Lin, X. S. (2012). Modeling dependent risks with multivariate Erlang mixtures. *ASTIN Bulletin*, 42(1):153–180.

Leung, K.-M., Elashoff, R. M., and Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual review of public health*, 18(1):83–104.

Li, Y., Gillespie, B. W., Shedden, K., and Gillespie, J. A. (2015). Calculating profile likelihood estimates of the correlation coefficient in the presence of left, right or interval censoring and missing data. *Working paper.*

Lin, T. I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 100(2):257 – 265.

Mailhot, M. (2012). *Mesures de risque et dépendance*. PhD thesis, Université Laval.

McLachlan, G. and Jones, P. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, pages 571–578.

McLachlan, G. and Peel, D. (2001). *Finite mixture models*. Wiley.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. Wiley-Interscience.

Nair, V. N. (1984). Confidence bands for survival functions with censored data: a comparative study. *Technometrics*, 26(3):265–275.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, 2nd edition.

Olsson, M. (1996). Estimation of phase-type distributions from censored data. *Scandinavian journal of statistics*, pages 443–460.

Peel, D. and McLachlan, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall.

Tijms, H. C. (1994). *Stochastic models: an algorithmic approach*. Wiley.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295.

Verbelen, R., Gong, L., Antonio, K., Badescu, A., and Lin, X. S. (2015). Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulletin*. To appear.

Willmot, G. E. and Lin, X. S. (2011). Risk modelling with the mixed Erlang distribution. *Applied Stochastic Models in Business and Industry*, 27(1):2–16.

Willmot, G. E. and Woo, J.-K. (2007). On the class of Erlang mixtures with risk theoretic applications. *North American Actuarial Journal*, 11(2):99–115.

Willmot, G. E. and Woo, J.-K. (2015). On some properties of a class of multivariate Erlang mixtures with insurance applications. *ASTIN Bulletin*, 45(01):151–173.

Zadeh, A. H. and Bilodeau, M. (2013). Fitting bivariate losses with phase-type distributions. *Scandinavian Actuarial Journal*, 2013(4):241–262.